Advanced IT Tools for Historical Research with Archival Material

Pavlos Fafalios

Centre for Cultural Informatics (CCI), Institute of Computer Science (ICS), FORTH

fafalios@ics.forth.gr

Outline

- Introduction
- The tools
  - **FastCat** (data transcription)
  - **FastCat Catalogues** (data browsing)
  - **FastCat Team** (data curation)
  - **SeaLiT Ontology** (data integration)
  - **SeaLiT ResearchSpace** (data exploration & analysis)
- Conclusion

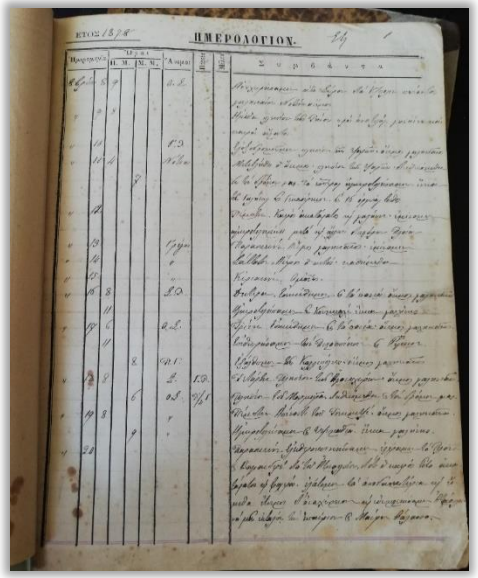
Introduction

Data Management in SeaLiT

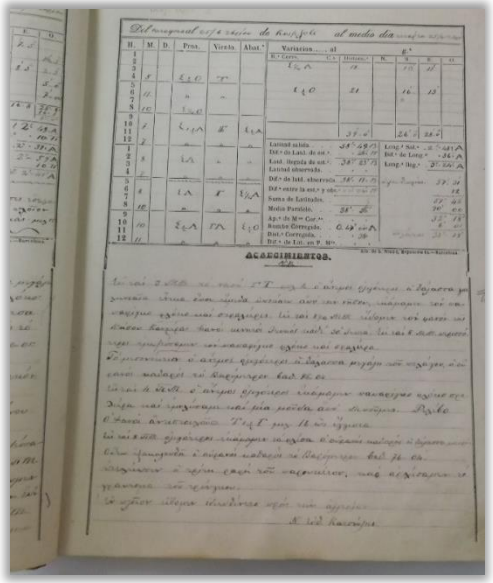
- The archival sources
  - Crew and displacement lists
  - Ship logbooks
  - Payrolls
  - Sailors registers
  - Naval Ship registers
  - Students registers
  - Employment records
  - Account books
  - Censuses
  - …

- *Provided by **different authorities** in several countries (**5 languages**)*
- *Containing **interconnected information** about different types of **entities** (sailors, ships, locations, …)*

Description of the full archival corpus:
https://sealitproject.eu/archival-corpus



Logbook (1878-1888)
Brig Eleni Koupa
Private Archive of Evangelos Rafalias, Hydra



Logbook (1882-1885)
Barquentine Asimoula
Maritime Museum of Galaxidi

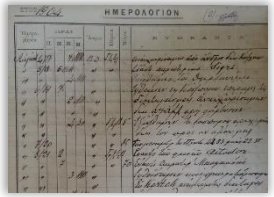
Data Management Challenges

- **Motivating Scenario**: Exploring information about a particular **ship**
  - **Problem 1:** Complementary information about the same ship may exist in different archival documents

Account book:
Information about
the **ship's owners**

Naval ship register list:
Construction details and
ship characteristics

Civil register:
Information about
ship's crew members



Crew and displacement list:
Ship's voyages and **crew members**

Need of integrating information coming from different archival documents!

Data Management Challenges

- **Motivating Scenario**: Exploring information about a particular **ship**
  - **Problem 2:** The **same ship**, or the same ship-related **entity/term,** might appear under different representations in the archival documents, due to:
    - Typos (“**Vindobonna**” vs “**Vindobona**”, “**Gaetano Schiaffino**” *vs* “**Gaetano Schiafino**”)
    - Different languages (“**Genoa**” *vs* “**Genova**”, “**brigantine**” *vs* “**brigantino**”)
    - Unrecognizable characters (“**G??oa**”)
    - Uncertainty (“**[Genoa]**”)
    - Use of abbreviation (“**Gaetano Schiaffino**” *vs* “**G. Schiaffino**”)

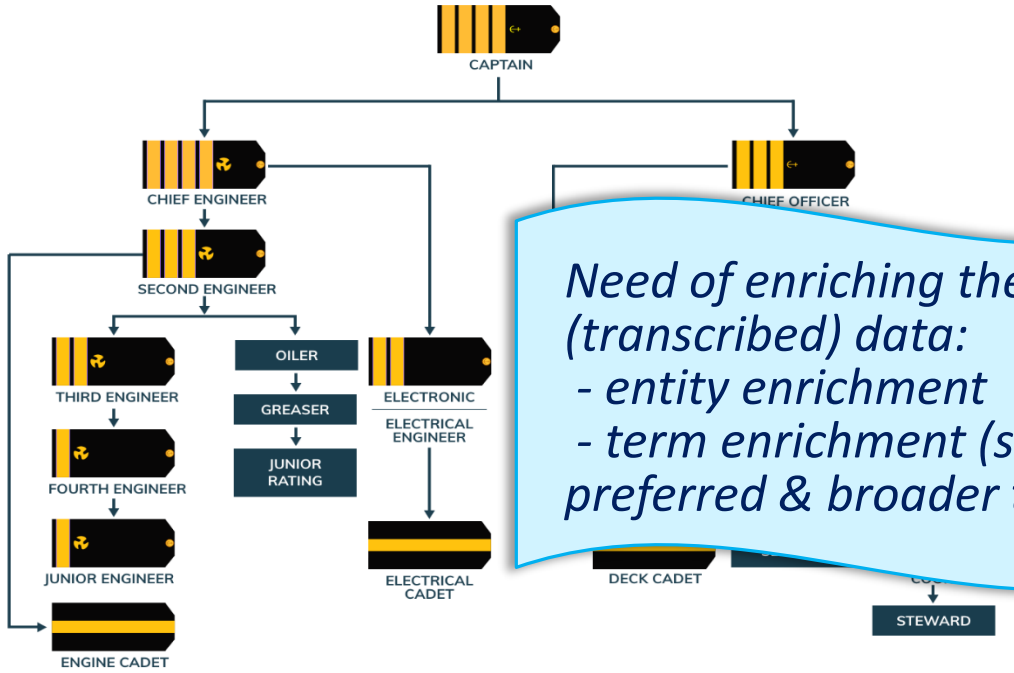
Need of curating the (transcribed) data:
- entity (instance) matching
- term alignment
- corrections

Data Management Challenges

- **Motivating Scenario**: Exploring information about a particular **ship**
  - **Problem 3:** Data **enrichment** is needed for better analysis and exploration services

Adding **coordinates** to locations
(for map visualization)

Creating **hierarchies** for terms
(e.g., for professions)



Need of enriching the (transcribed) data:
- entity enrichment
- term enrichment (setting preferred & broader terms)

Data Management Challenges

- **Motivating Scenario**: Exploring information about a particular **ship**
  - **Problem 4:** How to explore the (curated & integrated) data, both quantitatively and qualitatively, through user-friendly interfaces/visualizations?

    - Browsing the interconnected data

      *Starting from a specific **ship**, check its **arrival ports**, then find information about the **sailors** or other **ships** that arrived at that port …*

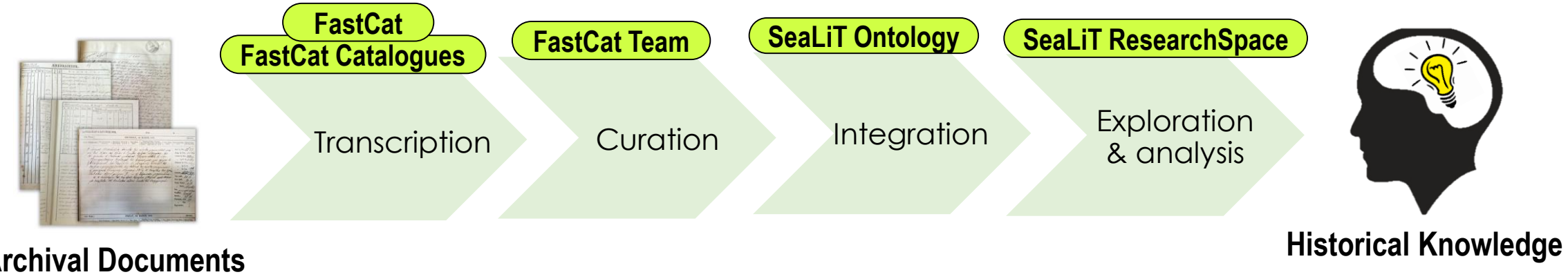
    - Running complex questions / aggregating information

      *"Find the **nationality** of **sailors** of **Greek ships** of a specific **type** (e.g., **Brig**) that arrived at a specific **port**"*

Need of tools/interfaces for supporting researchers in exploring and analyzing the (transcribed & curated) data

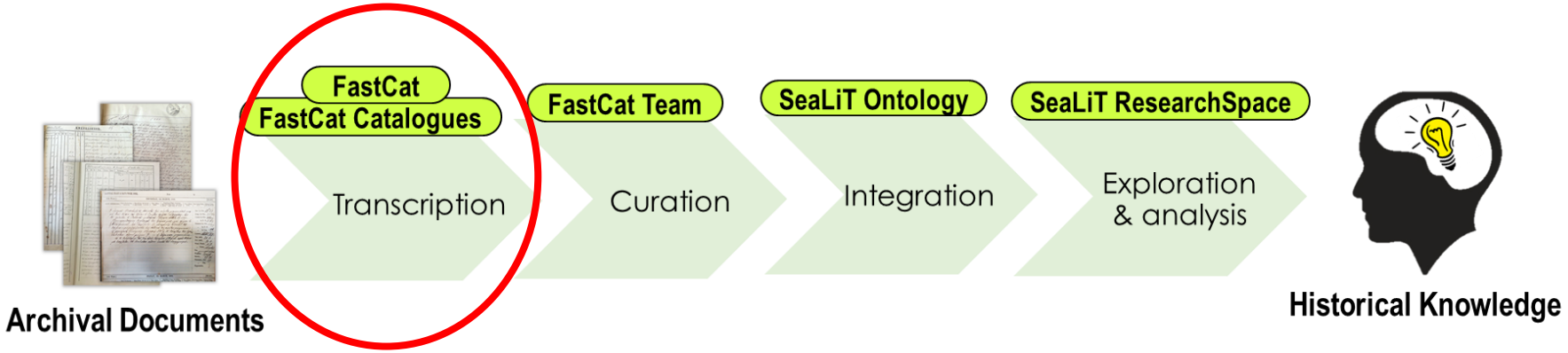
Our contribution

- **Workflow model** and **IT tools**



Focus on maintaining the important **data provenance** information!

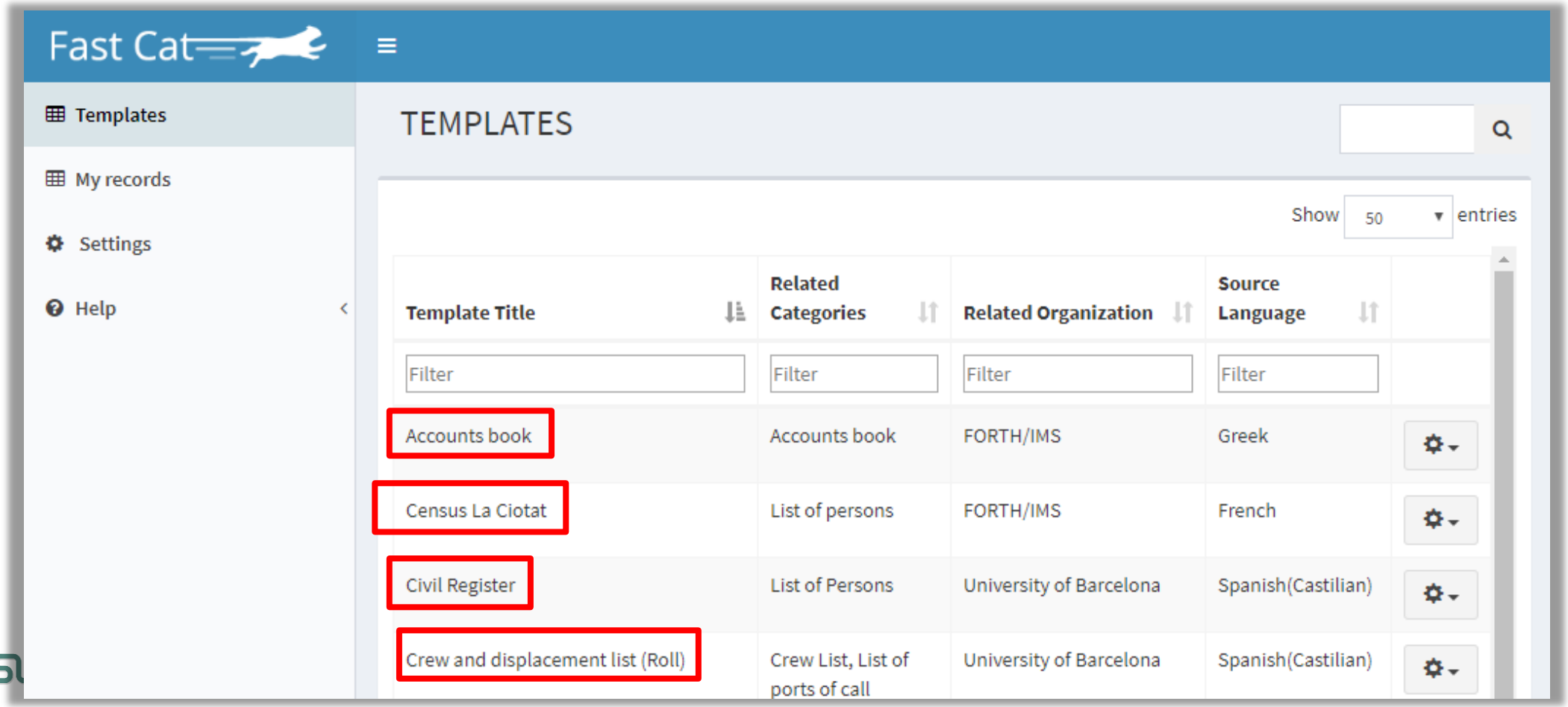
- Important for **verification** and the **long-term validity** of the research results



Data transcription with FastCat

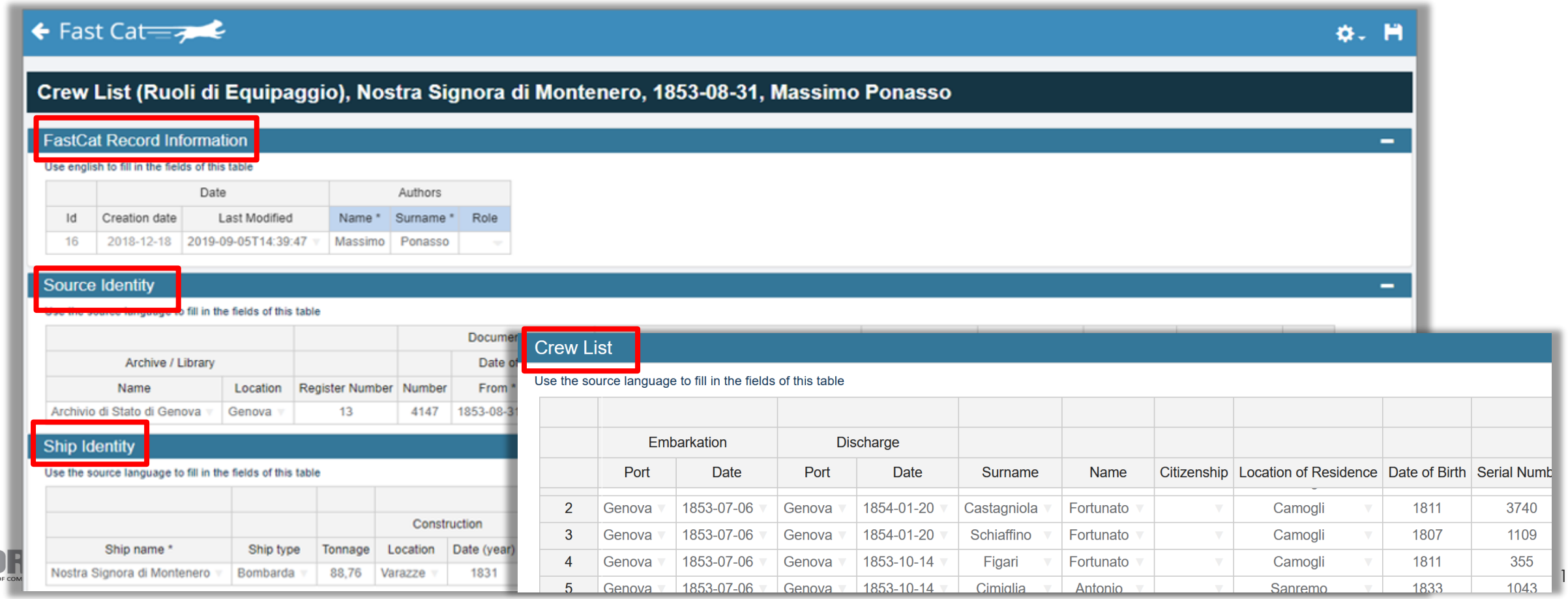
Data transcription with FastCat

- Data from different archival sources can be transcribed as '**records**' belonging to '**templates**'
  - A '**template**' represents the structure of a single data source



Data transcription with FastCat

- Data from different archival sources can be transcribed as '**records**' belonging to '**templates**'
  - A '**record**' organizes the data and metadata in tabular form (tables)



Data transcription with FastCat

- Use in **SeaLiT**:
  - ~30 users
  - 20 templates
  - >600 records
  - 5 languages

No	Template Name	# Records
1	Accounts book	14
2	Census La Ciotat	63
3	Civil Register	29
4	Crew and displacement list (Roll)	35
5	Crew List (Ruoli di Equipaggio)	98
6	Employment records, Shipyards of Messageries Maritimes, La Ciotat	50
7	First national all-Russian census of the Russian Empire	6
8	General Spanish Crew List	64
9	Maritime Register of the State for La Ciotat	1
10	List of ships	71
11	Logbook	17
12	Naval Ship Register List	2
13	Notarial Deeds	10
14	Payroll	7
15	Payroll of Russian Steam Navigation and Trading Company	14
16	Register of Maritime personel	4
17	Register of Maritime workers (Matricole della gente di mare)	6
18	Sailors register (Libro de registro de marineros)	52
19	Seagoing Personel	52
20	Students Register	10

Exploring the FastCat records with FastCat Catalogues

- FastCat Catalogues
  - A Web application for **browsing** and **analyzing (quantitively)** the data in all FastCat records
  - **Configurable** for use over other types of sources!

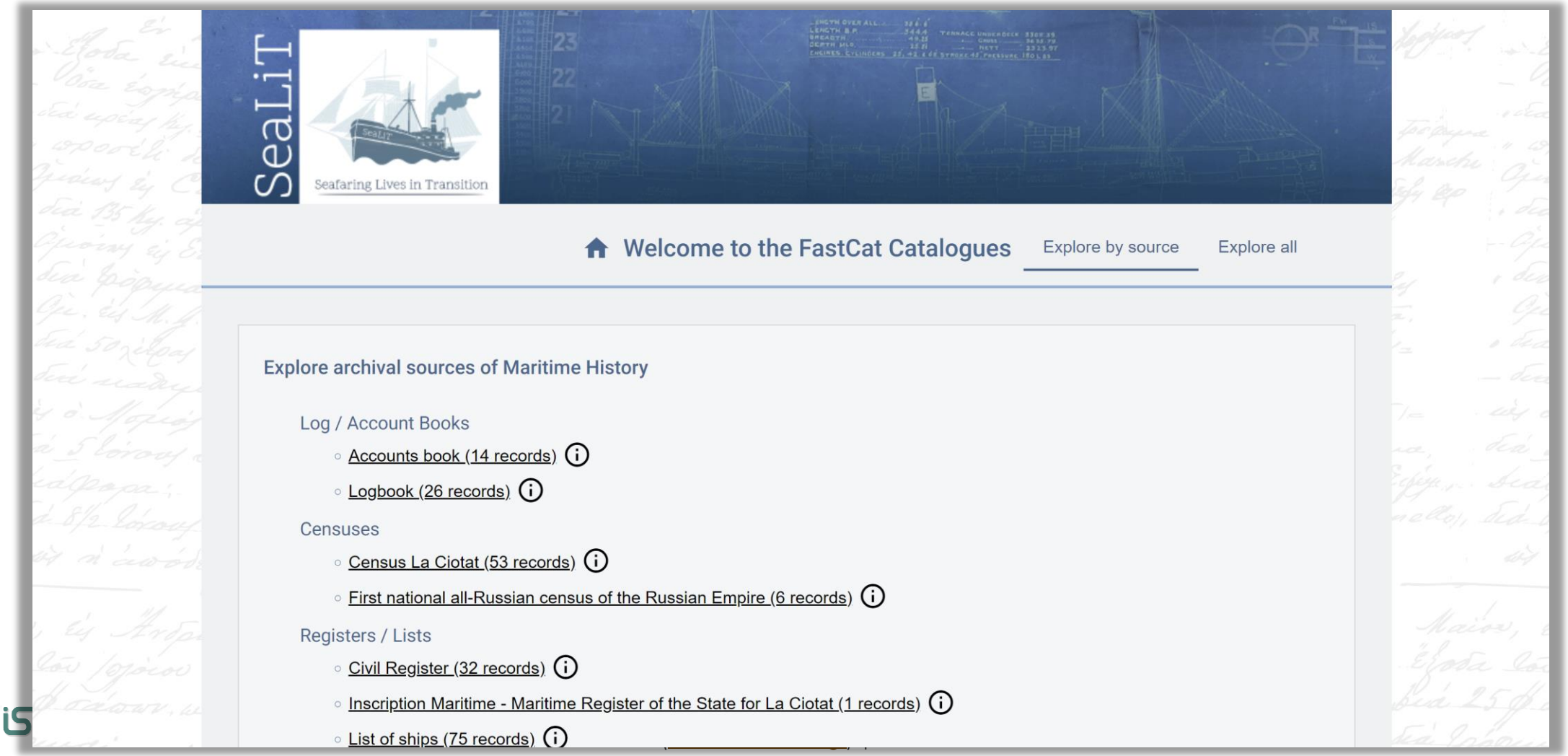
- Functionality:
  - **browsing** "entities of interest"(source-specific, source-independent)
  - data **ranking** and **filtering**
  - inspection of entity **provenance** information
  - data **aggregation** and **visualization** in **charts**
  - data **export** for further (external) analysis

https://catalogues.sealitproject.eu/

Available as **open source**:
https://github.com/isl/FastCat-Catalogues

FastCat Catalogues

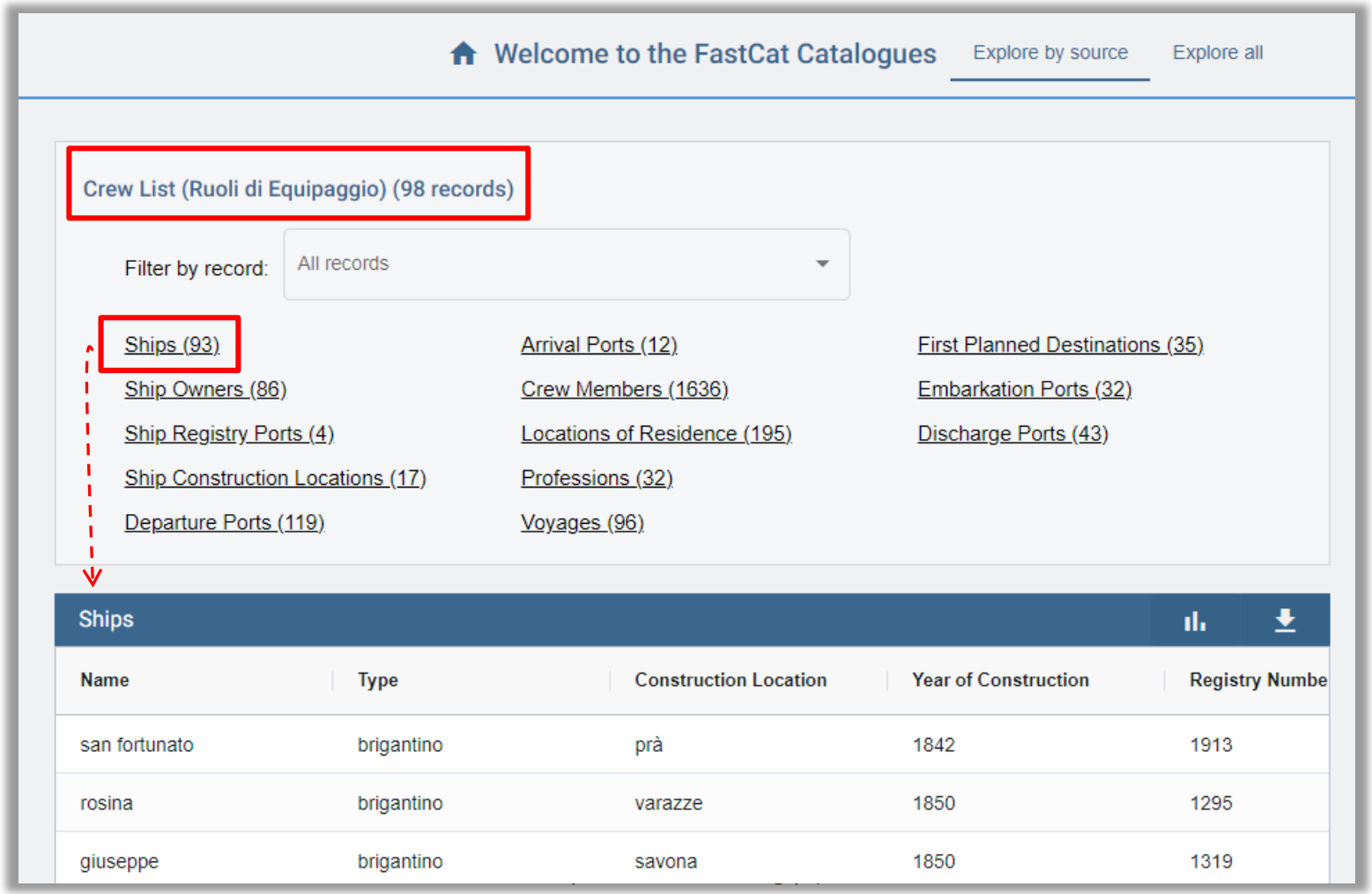
https://catalogues.sealitproject.eu/



FastCat Catalogues

https://catalogues.sealitproject.eu/

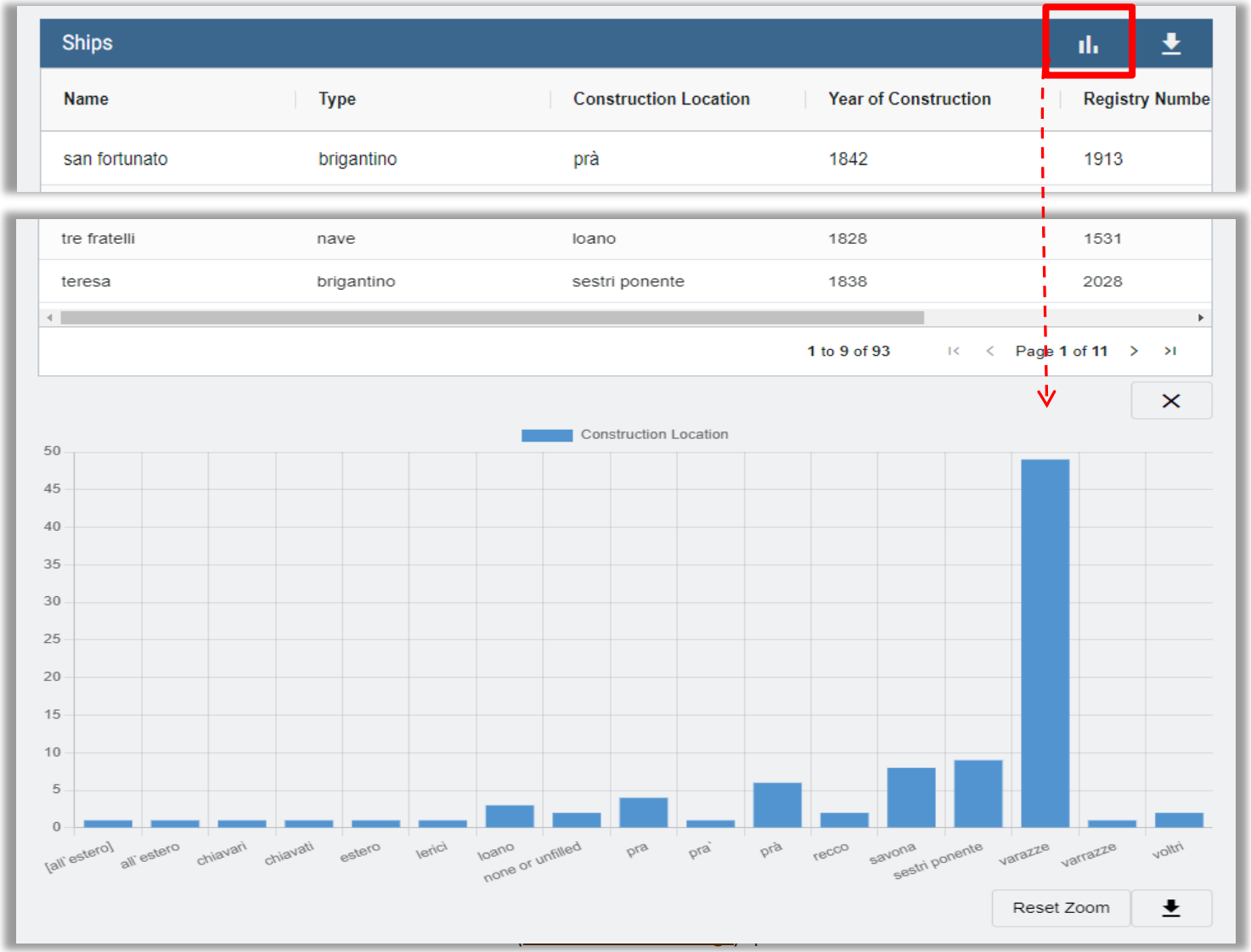
Data browsing and exploration

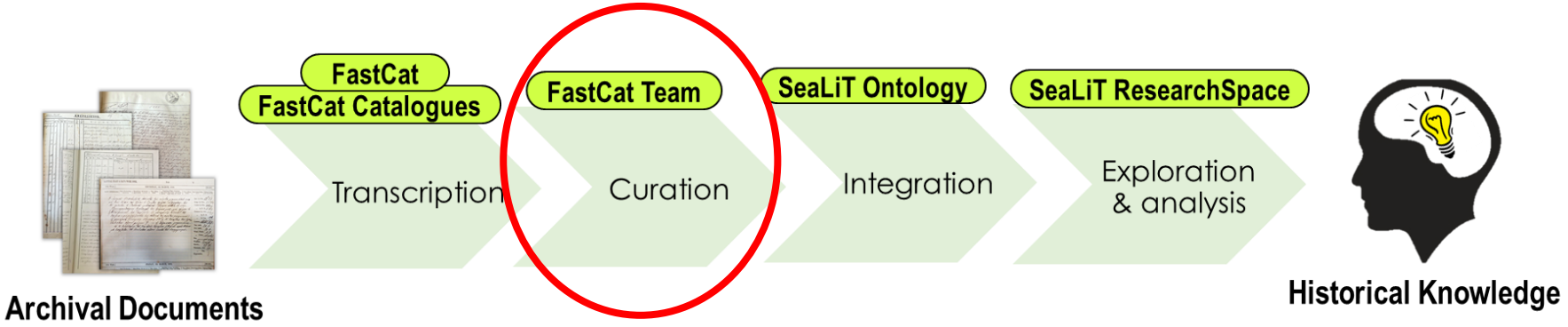


FastCat Catalogues

https://catalogues.sealitproject.eu/

Data visualization

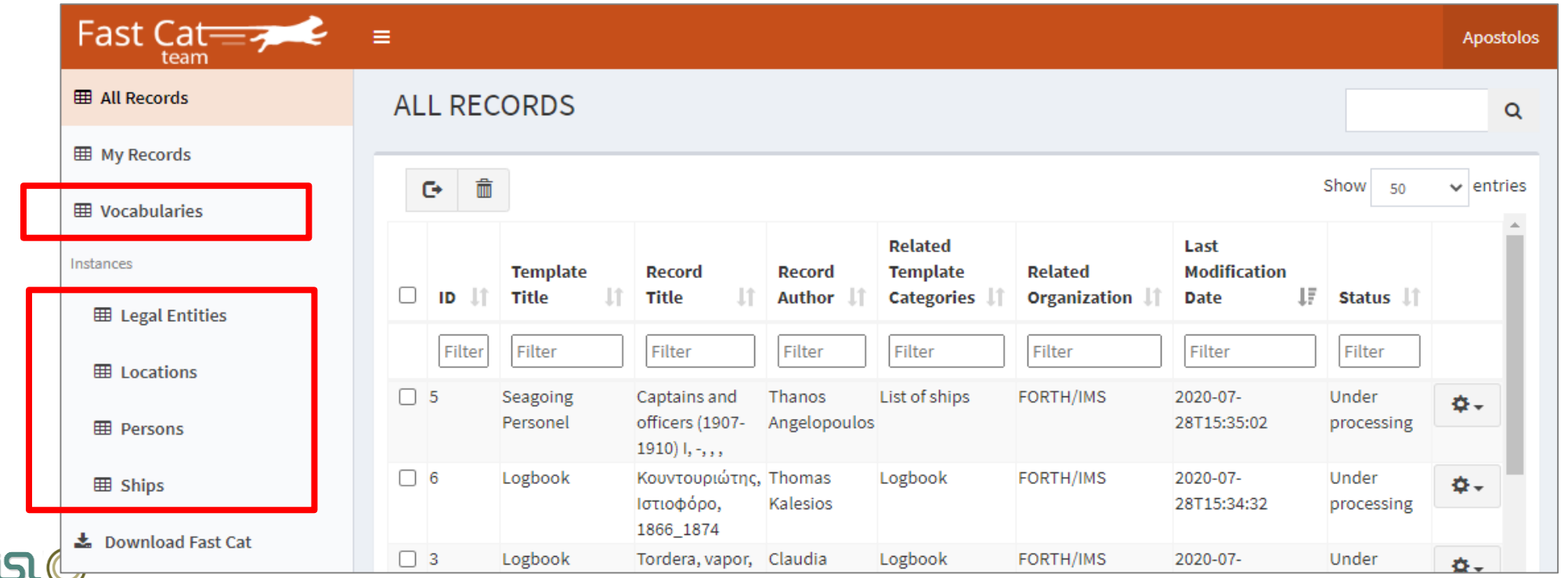




Data curation with FastCat Team

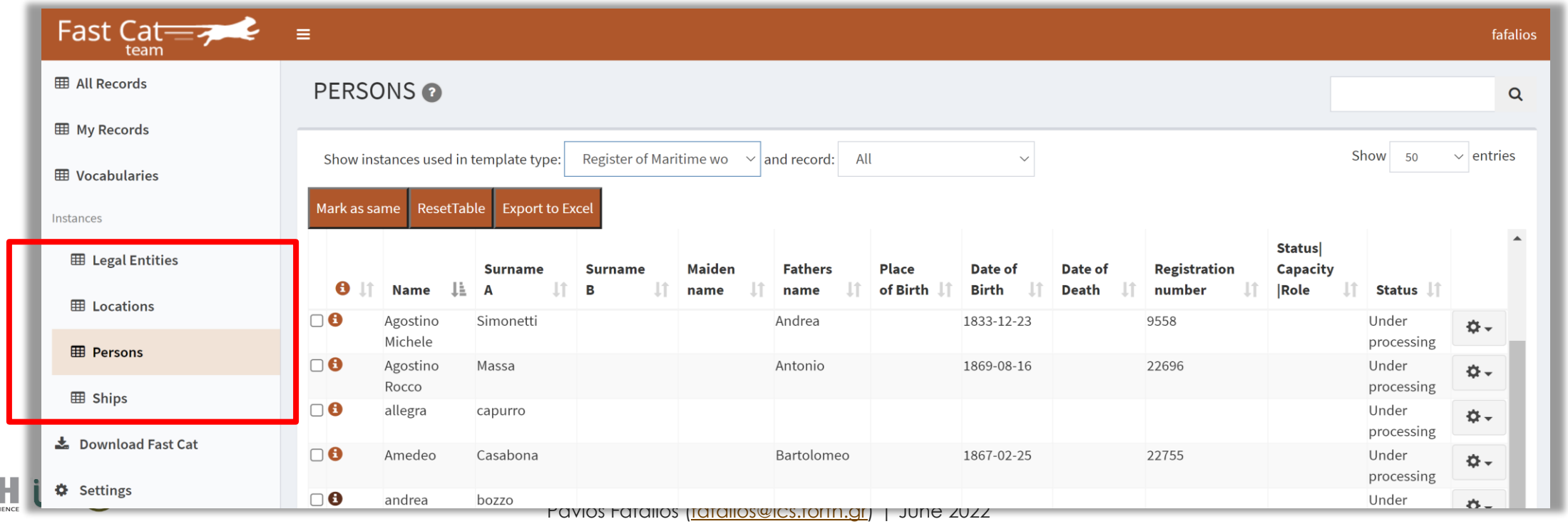
Data curation with FastCat Team

- **FastCat Team**:
  - A special environment within FastCat that allows the **collaborative curation** of the ***entities*** and ***vocabulary terms*** *that appear in the transcripts*
  - It **decouples** data curation from data transcription
    - We avoid 'spoiling' the original transcripts!



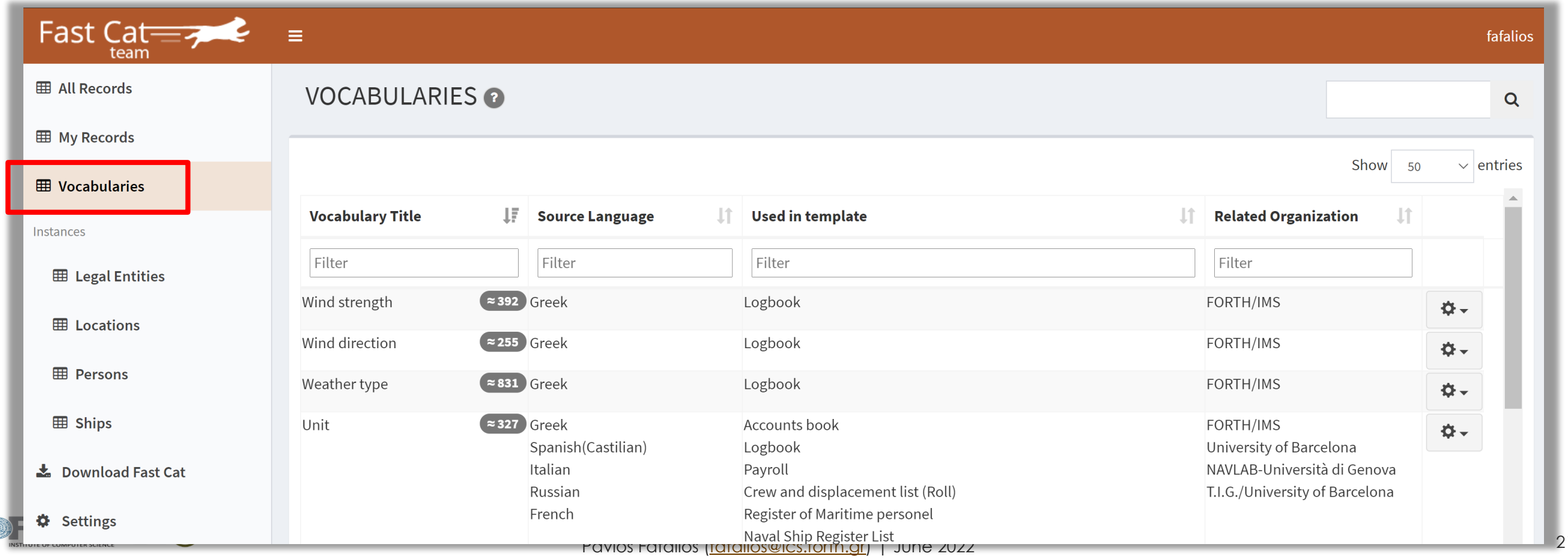
Data curation with FastCat Team

- Management of **entity instances**
  - Inspection of **records** in which the entity appears
  - **Correction** of entity names or other entity properties
  - Instance **matching** or **unmatching**
  - Entity **enrichment**



Data curation with FastCat Team

- Management of **vocabularies**
  - Inspection of **records** in which a vocabulary term appears
  - Setting a **preferred** and/or a **broader** term for a vocabulary term



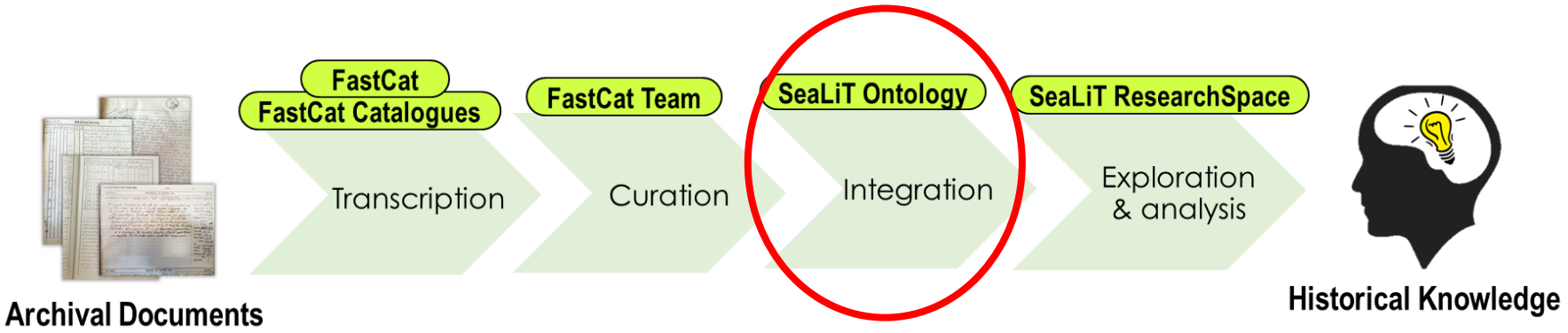
Data curation with FastCat Team

- Use in **SeaLiT**:
  - ~50 vocabularies *(ship type, profession, marital status, religion, …)*
  - >90K person instances
  - >9.8K location instances
  - >2.4K ship instances
  - >1.1K legal entity instances

FastCat & FastCat Team

- Available as open-source:
  - https://github.com/isl/FastCat

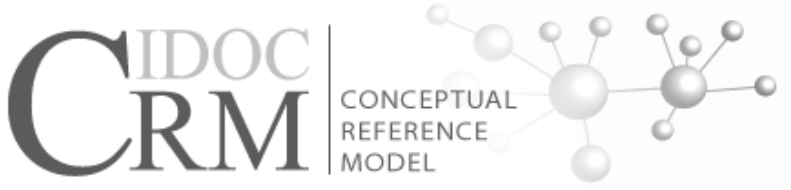
- Papers:
  - P. Fafalios et al. (2021), "FAST CAT: collaborative data entry and curation for semantic interoperability in digital humanities", *Journal on Computing and Cultural Heritage (JOCCH)*, *14*(4), 1-20. http://users.ics.forth.gr/~fafalios/files/pubs/fafaliosJOCCH2021.pdf
  - K. Petrakis et al. (2021), "Digitizing, Curating and Visualizing Archival Sources of Maritime History: the case of ship logbooks of the nineteenth and twentieth centuries". *Drassana, No. 28, 60–87.* https://doi.org/10.51829/Drassana.28.649



Data Integration with SeaLiT Ontology

The SeaLiT Ontology

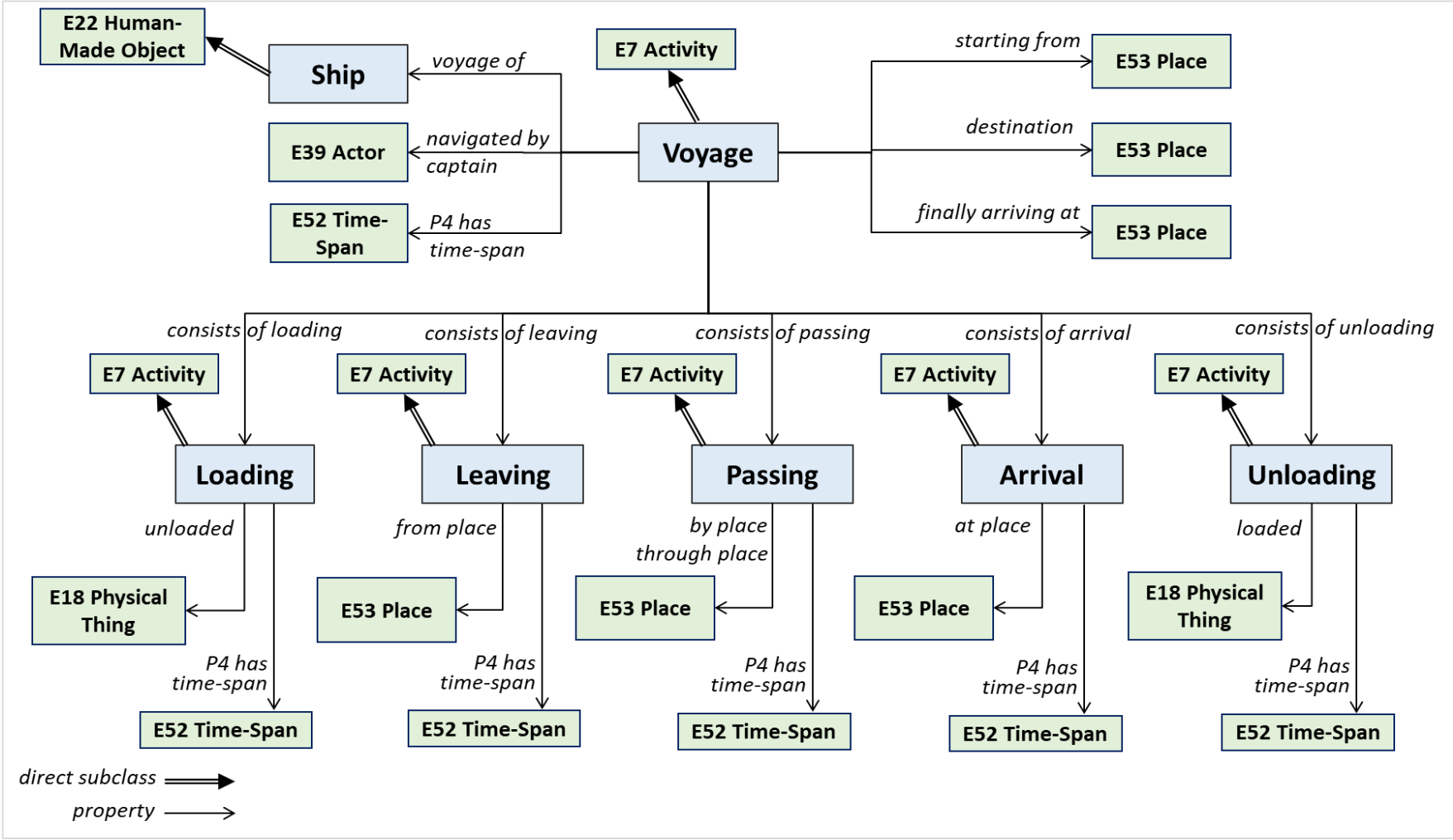
- It offers a **common** and **formal** '**language**' for describing information about Maritime History in a **structured** way
- Created as an extension of **CIDOC-CRM** (Conceptual Reference Model)
  - A high-level ontology (**ISO standard 21127:2014**) for describing concepts and relationships used in cultural heritage documentation (and beyond)



- Why using CIDOC-CRM?
  - It gives '**meaning**' to the data **(semantic interoperability**, **data sustainability)**
  - It facilitates **data integration** and **re-use**

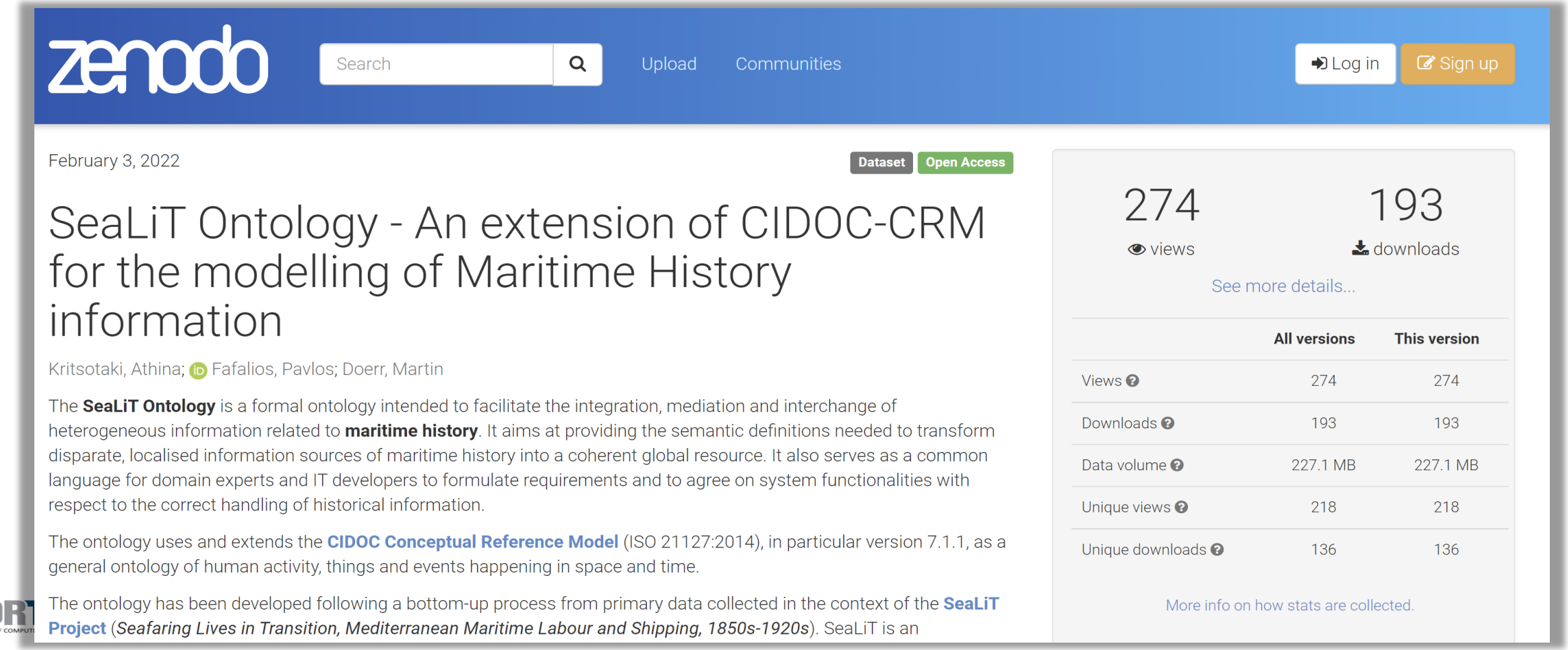
The SeaLiT Ontology

- A (small) part of the ontology



The SeaLiT Ontology

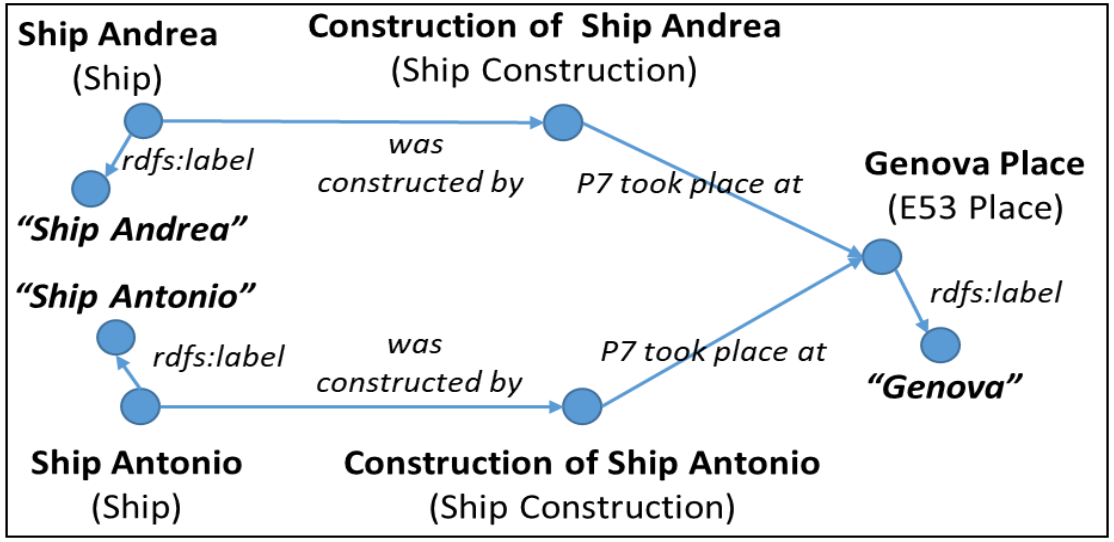
- It currently contains **46 classes** and **79 properties**
  - Available at: **https://zenodo.org/record/5964240**



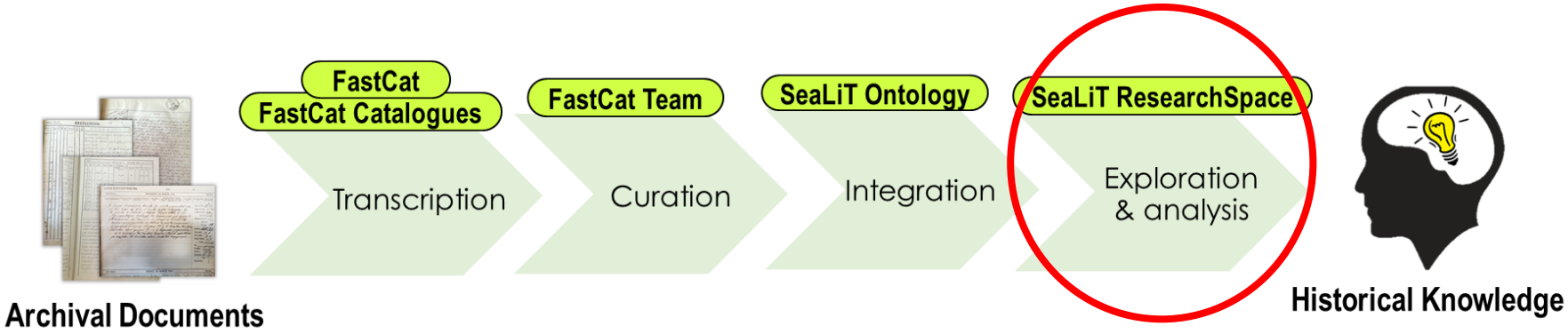
Creation of the Semantic Network

- Data transformation to a rich **semantic network** based on the **SeaLiT Ontology**
  - Using the **X3ML toolkit** (language, 3M user interface, transformation engine)
  - A (very small) part of the derived semantic network:

https://www.ics.forth.gr/isl/x3ml-toolkit



  - The full network contains >**18 million** edges!
    - Available at: **https://zenodo.org/record/6460841**



Data exploration and analysis with SeaLiT ResearchSpace

SeaLiT ResearchSpace

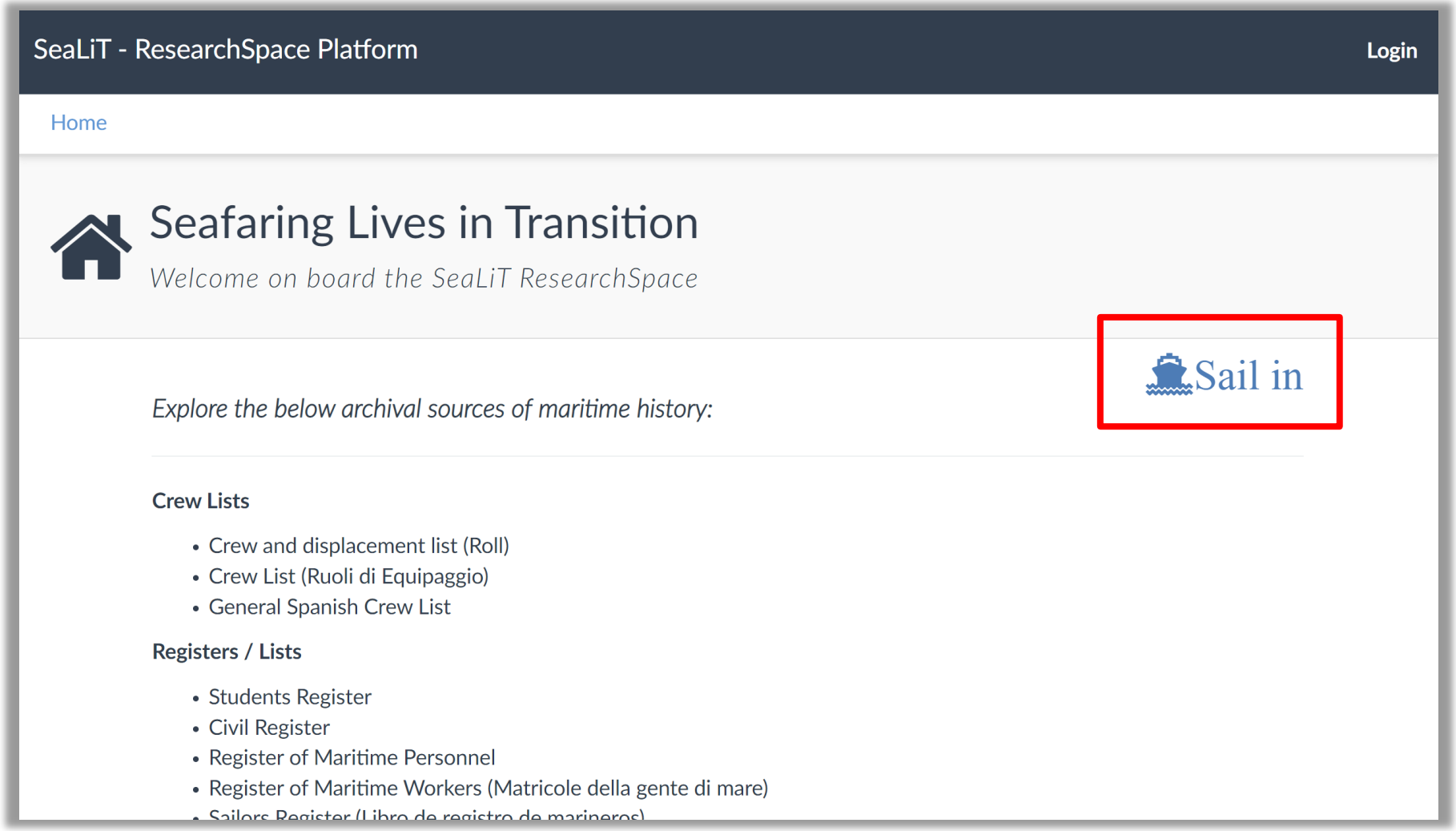
- **ResearchSpace**: open source platform, developed by the British Museum
- Offers a variety of functionalities:
  - Semantic search (through an assistive query building interface)
  - Data browsing
  - Variety of visualization methods (tables, charts, map)
  - Filtering (based on entity properties/relations)
- **Configured** for the case of the SeaLiT (integrated) data

https://www.researchspace.org/

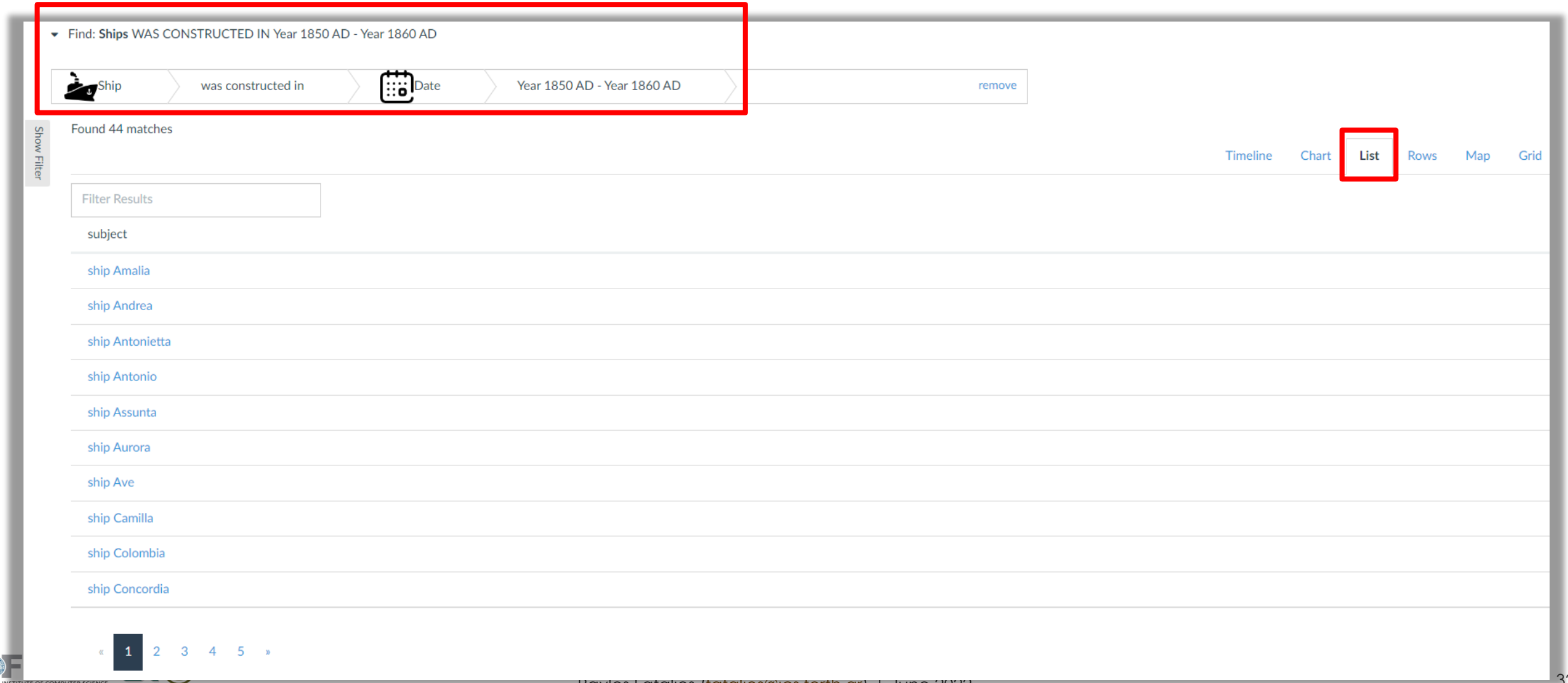
http://rs.sealitproject.eu/

SeaLiT ResearchSpace

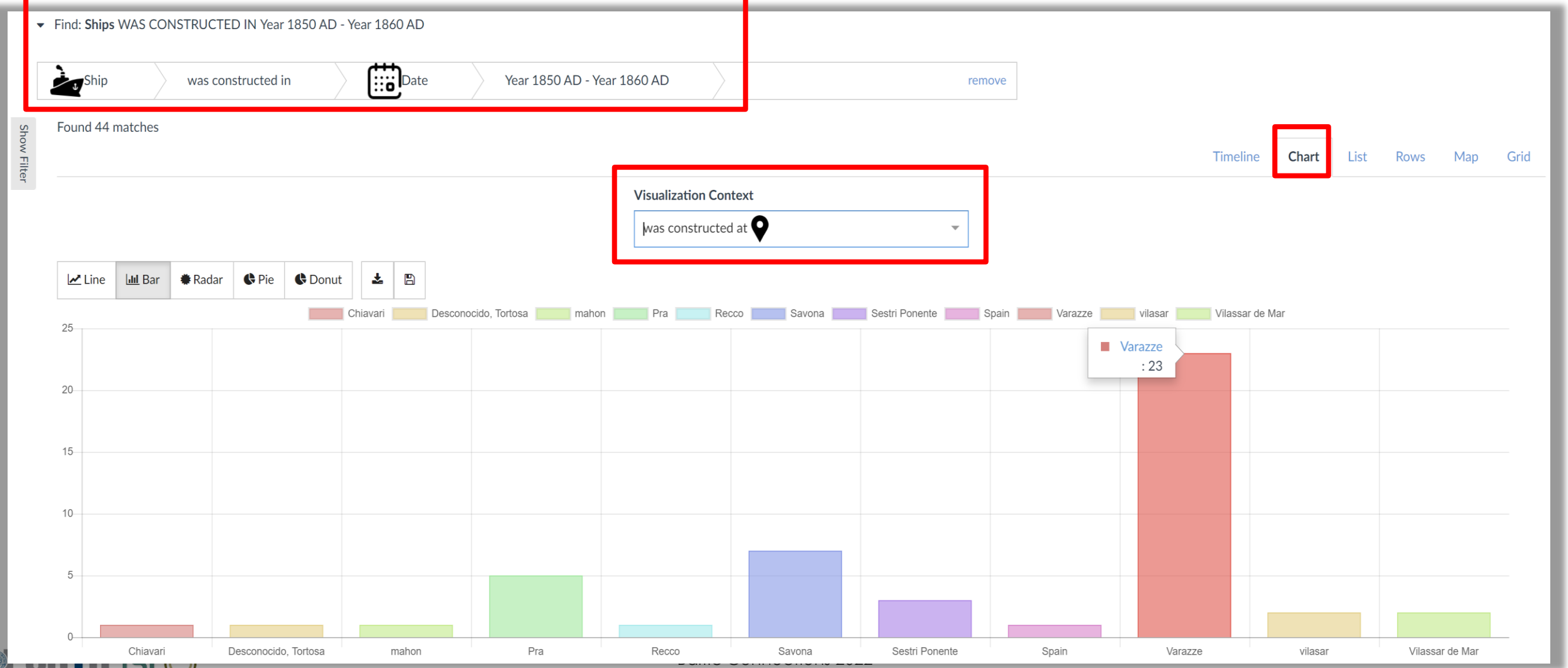
http://rs.sealitproject.eu/



Assistive query building in ResearchSpace



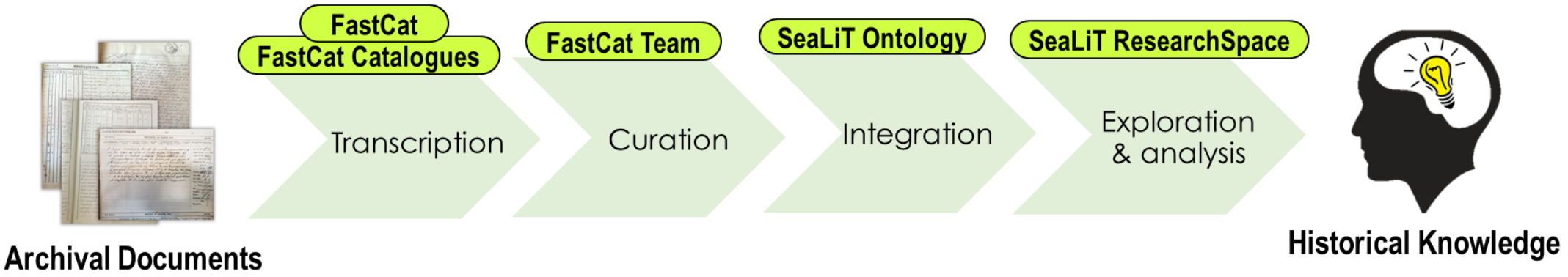
Assistive query building in ResearchSpace



Conclusion

Conclusion

- IT tools for <u>holistic data management</u> in historical-archival research



- **Data provenance-aware,** following established **documentation standards** (**CIDOC-CRM**)
- Successful **application** in a large scale research project of maritime history (**SeaLiT**)
- **Configurable** for use over other types of sources

Limitations / Future Work

- **Data entry errors** are common [FastCat]
  - Accidentally filling the wrong column, putting data in wrong place due to misunderstanding, …
  - **We need mechanisms that support users in detecting or avoiding such errors!**
- **Manual instance mapping** and **vocabulary curation** [FastCat Team]
  - Very **important**, but at the same time very **laborious** and **time consuming**
  - **We need tools that can automate these processes as much as possible, without significantly affecting quality!**
- **Comparing values** (**dates**, **quantities, etc.**) is sometimes difficult or impossible [FastCat / FastCat Team]
  - Due to use of different reference points or units across sources
  - **We need mechanisms to align such values!**

Try the data exploration applications:
https://sealitproject.eu/digital-seafaring

Thank you!

Pavlos Fafalios
fafalios@ics.forth.gr

The research & development team:

- Anastasia Axaridou (R&D engineer)
- Korina Doerr (Systems design engineer)
- Athina Kritsotaki (Data modeling engineer)
- Yannis Marketaks (R&D engineer)
- Kostas Petrakis (R&D engineer)
- Georgios Samaritakis (R&D engineer)
- Maria Theodoridou (R&D engineer)
- Pavlos Fafalios (Postdoctoral researcher)
- Martin Doerr (CCI Honorary Head)

Centre for Cultural Informatics (CCI)
Institute of Computer Science (ICS)
FORTH